ISSN: 2339-2541

JURNAL GAUSSIAN, Volume 5, Nomor 4, Tahun 2016, Halaman 757-762
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

PERBANDINGAN METODE *K-MEANS* DAN METODE DBSCAN PADA PENGELOMPOKAN RUMAH KOST MAHASISWA DI KELURAHAN TEMBALANG SEMARANG

Sisca Agustin Diani Budiman[1], Diah Safitri[2], Dwi Ispriyanti[3]
[1]Mahasiswa Departemen Statistika FSM Universitas Diponegoro
[2,3]Staff Pengajar Departemen Statistika FSM Universitas Diponegoro

ABSTRACT

Students as well as community or household, as well as economic activities daily, including consumption. The student needs to choose a place to stay is also one form of consumption activities. There are many factors that affect student preferences in the selection of boarding houses, including price, amenities, location, income, lifestyle, and others. The rental price boarding and facilities offered significant positive effect on student preferences in choosing a boarding house. Based on rental rates and facilities it offered to do the grouping in order to know the condition of the student boarding house in the Village Tembalang. Grouping is one of the main tasks in data mining and have been widely applied in various fields. The method used to classify is *K-Means* and DBSCAN with a number of groups of three. Furthermore, the results of both methods were compared using the Silhouette index values to determine which method is better to classify the student boarding house. Based on the research that has been conducted found that the K-Means method works better than DBSCAN to classify the student boarding house as evidenced by the value of the Silhouette index on *K-Means* of 0.463 is higher than the value at DBSCAN Silhouette index is equal to 0.281.

Keywords: student boarding houses, data mining, clustering, *K-Means*, DBSCAN

1. PENDAHULUAN

Mahasiswa sama seperti masyarakat atau rumah tangga, juga melakukan aktivitas ekonomi sehari-hari termasuk konsumsi. Kebutuhan mahasiswa untuk memilih tempat tinggal juga merupakan salah satu bentuk dari kegiatan konsumsi. Harga sewa kost dan fasilitas yang ditawarkan berpengaruh signifikan positif terhadap preferensi mahasiswa dalam memilih rumah kost[8]. Berdasarkan harga sewa dan fasilitas yang ditawarakan dapat dilakukan pengelompokan agar diketahui kondisi rumah kost mahasiswa di Kelurahan Tembalang.

Permasalahan yang dihadapi adalah data rumah kost mahasiswa merupakan kelompok data yang tidak tumpang tindih, sehinga metode pengelompokan yang digunakan untuk mengelompokan data rumah kost mahasiswa adalah metode pengelompokan sekatan (*partitioning*) dan eksklusif. *K-Means* dan DBSCAN merupakan metode pengelompokan sekatan dan eksklusif.

Metode *K-Means* adalah pengelompokan data nonhierarki (sekatan) yang mempartisi data ke dalam bentuk dua atau lebih kelompok, sehingga data yang berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama. *Density-Based Spatial Clustering Algorithm with noise* (DBSCAN) adalah algoritma pengelompokan yang didasarkan pada kepadatan (*density*) data. Konsep kepadatan dalam DBSCAN menghasilkan tiga macam status dari setiap data, yaitu inti (*core*), batas (*border*), dan *noise*.

Penelitian ini bertujuan untuk membandingkan kelompok mana yang paling optimal diantara kedua metode tersebut. Berdasarkan kelompok yang paling optimal kemudian dilakukan interpretasi adalah beberapa kelompok dengan interval harga tertentu dengan jumlah fasilitas yang didapatkan dari harga tersebut.

2. TINJAUANPUSTAKA

2.1 Data Mining

Data mining adalah proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Data mining merupakan teknologi yang menggabungkan metode analisis tradisional dengan algoritma yang canggih untuk memproses data dengan volume besar[10]. Data mining dapat digunakan untuk menyelesaikan masalah dalam bidang intelektual, ekonomi, dan bisnis dibagi menjadi enam tugas yaitu klasifikasi, estimasi, prediksi, afinitasi, *clustering*, deskripsi dan penetuan profil[2].

2.2 Pemrosesan Data Awal (*Preprocessing Data*)

Kumpulan data yang akan diproses dengan metode–metode dalam data mining sering kali harus melalui pekerjaan awal dikarenakan memiliki masalah seperti populasi data yang terlalu besar, dimensi data yang terlalu tinggi, banyaknya fitur yang tidak berkontribusi besar, dan adanya perbedaan skala pada variabel. Data *transformation* (transformasi data) adalah pekerjaan mengubah data ke dalam bentuk yang paling tepat atau cocok untuk proses data mining, untuk masalah data yang variabelnya memiliki skala yang berbeda dapat dilakukan normalisasi. Normalisasi min-max adalah melakukan transformasi linier pada data asli. Normalisasi ini mengubah nilai asli ke berbagai nilai data baru biasanya nilainya 0 – 1 dan dinyatakan dalam bentuk persamaan (1) [7].

$$v_i^{'} = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \quad (1)$$

2.3 Analisis Kelompok

Analisis Kelompok (*cluster analysis*) adalah pekerjaan mengelompokan data (objek) yang didasarkan hanya pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut dan hubungan diantaranya[10]. Dalam pengelompokan data, data yang akan dikelompokan tidak mempunyai label kelas, tetapi dikelompokan menurut karakteristiknya, barulah kelompok tersebut diberi label sesuai hasil karakteristik kelompok masing-masing. Karena alasan tersebut analisis kelompok sering disebut *segmentation* atau *partitioning*[7].

2.4 *K-Means*

K-Means merupakan salah satu metode pengelompokan data nonhierarki (sekatan) yang berusaha membagi data ke dalam bentuk dua atau lebih kelompok. Metode ini membagi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama dan data yang berbeda karakteristik dikelompokkan ke dalam kelompok yang lain[7]. Pengelompokan data dengan metode *K-Means* secara umum dilakukan dengan algoritma sebagai berikut[7] :

1. Menentukan banyaknya k kelompok.
2. Membagi data ke dalam k kelompok
3. Menghitung pusat kelompok (sentroid) dari data yang ada di masing-masing kelompok dan dinyatakan dalam bentuk persamaan (2).

$$C_i = \frac{1}{M}\sum_{j=1}^{M} x_j \quad (2)$$

   dimana C adalah sentroid, M adalah banyak data, i adalah banyak kelompok.
4. Masing-masing data dialokasikan ke sentroid terdekat. Menghitung jarak data ke setiap sentoid menggunakan jarak *Euclidean* dan dinyatakan dalam bentuk persamaan (3).

$$D(x_l, C_i) = \sqrt{\sum_{j=1}^{q}\left(x_{lj} - C_{ij}\right)^2} \quad (3)$$

5. Kembali ke langkah 3 apabila masih ada data yang berpindah kelompok

2.5 DBSCAN

Density – Based Spatial Clustering Algorithm With Noise (DBSCAN) adalah algoritma pengelompokan yang didasarkan pada kepadatan (*density*) data. Konsep kepadatan yang dimaksud dalam DBSCAN adalah banyaknya data (minPts) yang berada dalam radius Eps (ε) dari setiap data. Konsep kepadatan seperti ini menghasilkan tiga macam status dari setiap data, yaitu inti (*core*), batas (*border*), dan *noise*[7]. Data inti merupakan data yang jumlah data di dalam radius eps lebih dari minPts, data noise merupakan data yang jumlah data di dalam radius eps kurang dari minPts, dan data batas merupakan data yang jumlah data di dalam radius eps kurang dari minPts tetapi menjadikan data tetangganya menjadi data inti. Proses pengelompokan DBSCAN adalah menghitung jarak titik pusat (p) ke titik yang lain menggunakan jarak *Euclidean* dan dinyatakan dalam persamaan (4).

$$D(x_l, p_i) = \sqrt{\sum_{j=1}^{q}\left(x_{lj} - p_{ij}\right)^2} \quad (4)$$

Iterasi dilakukan pada setiap titik yang menjadi tetangga titik pusat (p) dan titik yang belum dikunjungi[3].

2.6 Validasi Kelompok

Hasil pengelompokan diuji tingkat validitasnya menggunakan Indeks Validitas Silhouette untuk menentukan jumlah cluster yang terbaik. Teknik ini memberikan representasi grafis singkat dari seberapa baik setiap objek terletak pada kelompok[3]. Analisa metode silhouette ini dilakukan dengan melihat besar nilai s. Hasil perhitungan nilai indeks validitas silhouette dapat bervariasi antara -1 hingga 1. Jika bernilai 1 maka objek berada dalam kelompok yang tepat, jika 0 maka objek tidak jelas harus masuk ke kelompok yang mana, dan jika -1 maka objek tidak tepat berada di kelompok tersebut[1]. Pengujian dilakukan dengan cara mencari nilai a(h) yaitu rata-rata jarak titik inti dengan semua titik pada kelompok yang sama, dan nilai b(h) yaitu rata-rata jarak titik inti dengan semua titik pada kelompok yang berbeda menggunakan jarak *Euclidean*. Setelah didapat nilai a(h) dan b(h), kemudian dicari nilai indeks *silhouette* dan dinyatakan dalam persamaan (5) [1].

$$S(h) = \frac{\left(b(h) - a(h)\right)}{\max(a(h), b(h))} \quad (5)$$

2.7 Sampel Acak Stratifikasi Proporsional

Sampel acak stratifikasi proporsional adalah jumlah proporsi masing – masing strata dalam sampel ditentukan secara proporsional sesuai dengan besarnya populasi. Strata terbesar akan mendapatkan sampel lebih besar dibandingkan dengan strata yang kecil. Stratifikasi proporsional baik digunakan ketika strata dalam populasi terbagi ke dalam tingkatan yang kurang lebih seimbang[4]. Untuk mendapatkan sampel yang dapat menggambarkan populasi, maka dalam penentuan ukuran sampel penelitian dapat digunakan rumus Slovin dan dinyatakan dalam persamaan (6).

$$u = \frac{N}{1 + Ne^2} \quad (6)$$

dimana u adalah ukuran sampel, N adalah ukuran populasi, dan e adalah persen kelonggaran ketidaktelitian karena kesalahan pengambilan sampel yang masih dapat ditolerir. Setelah diperoleh ukuran sampel populasi, kemudian dicari banyaknya sampel yang harus diambil pada masing – masing strata menggunakan stratifikasi proporsional dan dinyatakan dalam persamaan (7)[9].

$$u_a = u\left(\frac{N_a}{N}\right) \quad (7)$$

dimana u_a adalah sampel masing-masing strata, u adalah ukuran sampel, N_a adalah ukuran populasi masing-masing strata, dan N adalah ukuran populasi.

2.8 Rumah Kost

Rumah Kost adalah sebuah hunian digunakan oleh sebagian kelompok masyarakat sebagai tempat tinggal sementara atau hunian yang sengaja didirikan oleh pemilik untuk disewakan dengan sistem pembayaran per bulan[1].

2.9 Harga

Harga adalah sejumlah uang yang dibutuhkan untuk mendapatkan barang atau jasa. Harga ditentukan oleh banyaknya permintaan dan penawaran[5].

2.10 Fasilitas

Fasilitas yaitu segala sesuatu yang bersifat peralatan fisik dan disediakan oleh pihak penjual jasa untuk mendukung kenyamanan konsumen[6].

3. METODE PENELITIAN

3.1 Jenis dan Sumber Data

Data yang digunakan dalam Tugas Akhir ini merupakan data primer yang diperoleh melalui penyebaran kuesioner kepada responden yaitu rumah kost mahasiswa yang berada di Kelurahan Tembalang Semarang. Teknik sampling yang digunakan adalah *proportionate stratified random sampling*. Teknik sampling probabilitas dengan penentuan sampel berlapis berdasarkan proporsi populasi. Wilayah penelitian hanya pada Kelurahan Tembalang sebagai populasi, yang terdiri 8 RW sebagai strata, kemudian akan dilakukan penarikan sampel pada masing-masing strata secara proporsional. Variabel yang digunakan dalam penelitian ini adalah harga dan fasilitas.

3.2 Teknik Pengolahan Data

Teknik pengolahan data dalam penelitian ini adalah pengelompokan dengan metode K-Means dan metode DBSCAN menggunakan *software* R Statistic 3.3.1.

Tahap analisis data yang digunakan dalam penelitian ini sebagai berikut :

1. Membuat kuesioner.
2. Melakukan pengumpulan dan *preprocessing* data
3. Melakukan pengelompokan data rumah kost menggunakan metode K-Means dengan jumlah kelompok adalah tiga.
4. Melakukan pengelompokan data rumah kost menggunakan metode DBSCAN dengan epsilon = 0,1 dan minPts = 2.
5. Pengujian kelompok menggunakan uji *silhouette* pada kelompok yang telah dibentuk menggunakan metode K-Means dan DBSCAN
6. Melakukan analisis dari output yang dihasilkan.

4. HASIL DAN PEMBAHASAN

4.1 Analisis Deskriptif

Populasi rumah kost di Kelurahan Tembalang sebanyak 1200 rumah, setelah dilakukan perhitungan dengan menggunakan rumus Slovin dihasilkan sampel sebanyak 300 rumah. Dilakukan perhitungan untuk mencari banyaknya sampel pada masing-masing strata dan diperoleh 50 sampel untuk RW 1, RW 2, RW 3, RW 4, RW 7, 45 sampel untuk RW 5, dan 5 sampel untuk RW 8.

4.2 Pengelompokan dengan Metode *K-Means*

Hasil pengelompokan data rumah kost menggunakan metode *K-Means* dengan jumlah kelompok tiga tersaji dalam Tabel 1.

Tabel 1. Hasil Pengelompokan *K-Means*

Kelompok	1	2	3
Jumlah Data	24	145	131

4.3 Pengelompokan dengan Metode DBSCAN

Hasil pengelompokan data rumah kost menggunakan metode DBSCAN dengan nilai eps adalah 0,1 dan minPts adalah 2 tersaji dalam Tabel 2.

Tabel 2. Hasil Pengelompokan DBSCAN

Kelompok	1	2	3
Jumlah Data	292	2	4

Terdapat dua data menjadi data noise karena tidak memenuhi syarat eps dan minPts.

4.4 Perhitungan Nilai Indeks *Silhouette*

Hasil perhitungan nilai indeks *Silhouette* pada kelompok yang dihasilkan dari pengelompokan menggunakan *K-Means* dan DBSCAN tersaji dalam Tabel 3.

Tabel 3. Indeks *silhouette* pada *K-Means* dan DBSCAN

Nilai	*K-Means*	DBSCAN
Indeks Silhouette	0,463	0,281

Dari Tabel 3, dapat disimpulkan bahwa pengelompokan menggunakan *K-Means* lebih baik dari DBSCAN, karena *K-Means* menghasilkan nilai indeks *Silhouette* lebih tinggi.

4.5 Interpretasi Hasil Pengelompokan yang Paling Optimal

Berdasarkan hasil perhitungan nilai Indeks Silhouette diketahui bahwa pengelompokan menggunakan metode *K-Means* lebih baik dibandingkan dengan metode DBSCAN, sehingga interpretasi kelompok yang dihasilkan menggunakan metode *K-Means* disajikan dalam Tabel 4.

Tabel 4. Interpretasi variabel masing-masing kelompok

Variabel	Kelompok 1	Kelompok 2	Kelompok 3
Harga	0,68515	0,205486	0,262354
Fasilitas	0,777778	0,347701	0,700382

Berdasarkan Tabel 4, dilakukan perhitungan min-max untuk mendapatkan harga dan fasilitas masing-masing kelompok. Harga dan fasilitas masing-masing kelompok disajikan dalam Tabel 5.

Tabel 5. Harga dan fasilitas masing-masing kelompok

Variabel	Kelompok 1	Kelompok 2	Kelompok 3
Harga	1.111.250	473.297	548.932
Fasilitas	13	8	11

5. KESIMPULAN

Pengelompokan data rumah kost menggunakan metode *K-Means* dengan jumlah kelompok tiga menghasilkan kelompok 1 dengan 24 data, kelompok 2 dengan 145 data, dan kelompok 3 dengan 131 data. Sedangkan pengelompokan menggunakan metode DBSCAN dengan nilai eps adalah 0,1 dan minPts adalah 2 menghasilkan kelompok 1 dengan 292 data, kelompok 2 dengan 2 data, kelompok 3 dengan 4 data, dan 2 data menjadi data noise.

Hasil perhitungan indeks *silhouette* menunjukan metode *K-Means* menghasilkan nilai 0,463 dan metode DBSCAN menghasilkan nilai 0,281, sehingga metode *K-Means* lebih baik dari metode DBSCAN dalam mengelompokan data rumah kost.

Hasil interpretasi kelompok optimal yaitu kelompok yang dibentuk menggunakan metode *K-Means* yaitu kelompok 1 dengan harga 1.111.250 terdapat 13 fasilitas, kelompok 2 dengan harga 473.297 terdapat 8 fasilitas, dan kelompok 3 dengan harga 548.932 terdapat 11 fasilitas.

6. DAFTAR PUSTAKA

[1] Alfina, T., Santosa, B. & Barakbah, A.R. 2012. *Analisa Perbandingan Metode Hierarchical Clustering, K-Means dan Gabungan Keduanya dalam Cluster Data*. JURNAL TEKNIK ITS, 1, pp.A521-25.

[2] Berry, M.J.A. & Linoff, G.S. 2004. *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management Second Edition*. Indianapolis, Indiana: Wiley Publishing, Inc.

[3] Devi, N.M.A.S., Putra, I.K.G.D. & Sukarsa, I.M. 2015. *Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan*. Lontar Komputer, 6 No.3.

[4] Eriyanto. 2007. *Teknik Sampling Analisis Opini Publik*. Yogyakarta: LKis Yogyakarta.

[5] Gitosudarmo, I. 2008. *Manajemen Pemasaran*. Kedua ed. Yogyakarta: BPFE.

[6] Kotler, P. & Keller, K.L. 2005. *Manajemen Pemasaran Jilid II*. Jakarta: PT Indeks Kelompok Gramedia.

[7] Prasetyo, E. 2012. *Data Mining Konsep dan Aplikasi menggunakan MATLAB*. Yogyakarta: Andi.

[8] Rachmawati, S. 2013. Analisis Preferensi Mahasiswa dalam Pemilihan Tempat Kos. *Jurnal Ilmiah Mahasiswa FEB Universitas Brawijaya*, p.vol 2 No 1.

[9] Scheaffer, R.L., Mendenhall, W. & Ott, L. 1990. *Elementary Survey Sampling Fourth Edition*. Boston: PWS-KENT Publishing Company.

[10] Tan, P.N., Steinbach, M. & Kumar, V. 2006. *Introduction to Data Mining*. New York: Pearson Addison Wesley.